HOW MANY EVOLUTIONARY HISTORIES ONLY INCREASE FITNESS?

JULIEN BERESTYCKI, ÉRIC BRUNET, AND ZHAN SHI

Abstract. Motivated by an evolutionary biology question, we study the following problem: we consider the hypercube $\{0,1\}^L$ where each node carries an independent random variable uniformly distributed on $[0,1]$, except $(1,1,\ldots,1)$ which carries the value 1 and $(0,0,\ldots,0)$ which carries the value $x \in [0,1]$. We study the number θ of paths from the root $(0,0,\ldots,0)$ to the opposite corner $(1,1,\ldots,1)$ along which the values on the nodes form an increasing sequence. We show that if the value on the root is set to $x = X/L$ then θ/L converges in law as $L \to \infty$ to e^{-X} times the product of two standard independent exponential variables.

As a first step in the analysis we study the same question when the graph is that of a tree where the root has arity L, each node at level 1 has arity $L-1$, $\ldots$, and the nodes at level $L-1$ have only one offspring which are the leaves of the tree (all the leaves are assigned the value 1, the root the value $x \in [0,1]$).

1. Introduction

We consider the following very simplified model for the evolution of an organism. The genetic information of the organism is encoded into its genome which, for our purposes is a chain of L_0 sites. With time, the organism accumulates mutations.

If we suppose that there are only two possible alleles on each site, it makes sense when looking at the genome to only record whether the allele carried at a given site is the original one or the mutant. We will represent a genetic type by a sequence of 0's and 1's of length L_0 where we put a 0 at position i if this site carries the original code or a 1 if it carries the mutant. Hence, a genetic type is a point $\sigma \in \{0,1\}^{L_0}$, the L_0-dimensional hypercube.

As an organism evolves by successive mutations, its genetic type travels along the edges of the hypercube. We suppose that each genetic type $\sigma \in \{0,1\}^{L_0}$ is characterized by a certain fitness value x_σ. We assume that the population is in a regime with a low mutation rate and strong selection; this means that when a new genetic type (mutant) appears in a resident population one of two things can happen: either the mutant type has better fitness and it fixates (i.e. it invades the whole population and becomes the resident type), or its fitness is lower and it becomes extinct (i.e. no one in the population carries this type after some time). Therefore, in that low mutation and strong selection regime, the only possible evolutionary paths are such that the fitness is always increasing. We say that such paths are *open*.

Somewhere in the L_0-dimensional hypercube, there is a type with the highest fitness. We call L the distance between that type and the original one; i.e. L is the number of

2000 *Mathematics Subject Classification.* Primary 60J80; Secondary 92D15.

Key words and phrases. Evolutionary biology, percolation, trees, branching processes.

mutated alleles in the fittest type. A natural question is whether there is an open path from the original type to the fittest. Such a path has at least L steps but may contain many more. Because of the low mutation rate, evolution takes time, and we interest ourselves here only in the *shortest* open paths leading to the fittest type, that is in the paths with exactly L steps, for which mutation never goes back: a site can only change from the original type to the mutant one.

In that setting, it is sufficient to consider the L-dimensional hypercube which contains (as opposing nodes) the original type, noted $\sigma_0 = (0, 0, \dots, 0)$ and the fittest type $\sigma_L = (1, 1, 1, \dots, 1)$. We consider paths through that hypercube along the edges which always move further away from the origin (i.e. at each step a 0 is changed into a 1 in the sequence) and, out of the $L!$ possible paths, we wish to count the number θ of open paths, that is the number of paths such that the fitness values form an increasing sequence. In biology, paths with increasing fitness values are also referred to as selectively accessible (see [8], [7], [1]).

To count the number of open paths, we need to know the fitness values of all the nodes. As a zero-model, we consider the case where the fitness values are random variables chosen independently with a common distribution, the so called "House of Cards" model ([4], [3]). As we are only interested in whether a sequence is increasing or not, the results will not depend on the specific distribution (as long as it is atomless). We therefore choose to give a fitness 1 to the fittest node and to assign uniform random numbers between 0 and 1 to each other nodes.

To summarize, motivated by this simple model, the problem we consider is the following: given an L-dimensional hypercube with L large, how many oriented paths are there from $\sigma_0 = (0, 0, \dots, 0)$ to $\sigma_L = (1, 1, \dots, 1)$:

$$\sigma_0 \to \sigma_1 \to \sigma_2 \to \cdots \to \sigma_L,$$

where each σ_{i+1} is obtained from σ_i by changing a single 0 into a 1 in the sequence σ_i, such that the fitness values x_σ form an increasing sequences:

$$x_{\sigma_0} < x_{\sigma_1} < \cdots < x_{\sigma_L},$$

where $x_{\sigma_L} = 1$ and where all the other x_σ are assigned random independent fitness values uniformly chosen in $[0, 1]$? A variant to this model which we also consider is when the starting point is assigned a non-random fitness value $x_{\sigma_0} = x$ with x given.

The correlation structure of the hypercube raises significant technical challenges. As a first step, we study the following simplified problem: instead of working on the L-dimensional hypercube we chose to work on a deterministic rooted tree as in Figure 1 with arity decreasing from L to 1: the root is connected to L first level nodes, each first level node is connected to $L - 1$ second level nodes, etc. There are L levels in the tree and $L!$ directed paths. The number of possible steps at level k is then $L - k$, as on the hypercube. Each of the $L!$ leaves of the tree (at level L) are assigned the value 1. All the other nodes are assigned independent random numbers uniformly drawn between 0 and 1, except perhaps the root to which we may choose to give a fixed value x. We are interested in the number θ of directed paths on the tree going from the root to one of the leaves where the numbers assigned to the visited nodes form an increasing sequence. As before, such a path is said to be open.

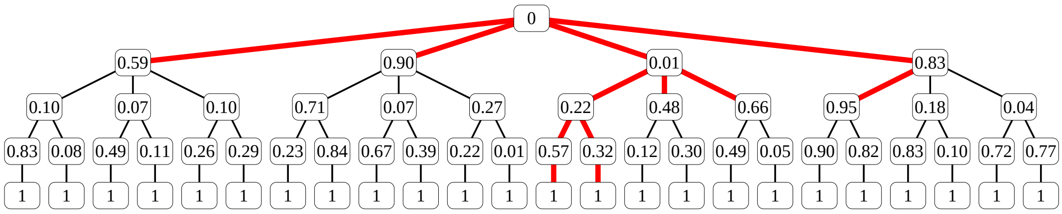

FIGURE 1. A tree for $L = 4$ and $x = 0$. The bold red lines indicate the directed paths going down from the root (at the top) and which visit an increasing sequence of numbers. There are $\theta = 2$ paths going all the way down to the leaves of the tree.

We mention that other models of paths with increasing fitness values on trees have also been considered in the literature [6].

2. Main results

Throughout the paper we use the following notations for the probability of an event and for the expectation and the variance of a number:

$$\mathbb{P}^x(\cdot), \quad \mathbb{E}^x(\cdot), \quad \mathrm{Var}^x(\cdot) \qquad \text{when the root has value } x$$

$$\mathbb{P}^*(\cdot) = \int_0^1 \mathrm{d}x\, \mathbb{P}^x(\cdot), \quad \mathbb{E}^*(\cdot) = \int_0^1 \mathrm{d}x\, \mathbb{E}^x(\cdot), \quad \mathrm{Var}^*(\cdot) \quad \text{when the root is uniform in } [0,1]$$

Note that the size L of the system is implicit in the notation. Whether we work on the tree or on the hypercube will always be made clear from the context. We call θ the number of open paths. The notation $a_L \sim b_L$, $L \to \infty$ means $\lim_{L\to\infty} \frac{a_L}{b_L} = 1$.

Obtaining the expectation of the number of paths for a given x is easy: there are $L!$ paths in the tree or the hypercube. Each path has probability $(1-x)^{L-1}$ that the $L-1$ intermediate numbers between the root and the leaf are between x and 1. Furthermore, there is probability $1/(L-1)!$ that these intermediate numbers form an increasing sequence. Hence, the probability that a given path is open is $(1-x)^{L-1}/(L-1)!$ and thus

$$\mathbb{E}^x(\theta) = L(1-x)^{L-1} \tag{2.1}$$

both for the tree and the hypercube.

Thus, if $x = 0$ there are on average L open paths, and if $x > 0$ the number of open paths goes in probability to zero when L becomes large.

The most biologically relevant variant of the model is when x is also randomly picked as the other nodes. The expectation of θ (both on the tree and on the hypercube) is trivially obtained by integrating (2.1):

$$\mathbb{E}^*(\theta) = 1, \tag{2.2}$$

but the typical number of paths for L large is not of order 1, as can be seen by looking on the variance of θ:

$$\lim_{L\to\infty} \frac{\mathrm{Var}^*(\theta)}{L} = 1 \qquad \text{on the tree.} \tag{2.3}$$

(All the variance computations on the tree are carried out in section 5.) In fact, this can be understood by considering starting values x scaling with the size L of the system as $x = X/L$ with X fixed:

Proposition 1. *In the case of the tree,*

$$\lim_{L\to\infty} \mathbb{E}^{\frac{X}{L}}(\theta/L) = e^{-X}, \qquad \lim_{L\to\infty} \mathrm{Var}^{\frac{X}{L}}(\theta/L) = e^{-2X}. \tag{2.4}$$

In the case of the hypercube (Hegarty-Martinsson [2]*)*

$$\lim_{L\to\infty} \mathbb{E}^{\frac{X}{L}}(\theta/L) = e^{-X}, \qquad \lim_{L\to\infty} \mathrm{Var}^{\frac{X}{L}}(\theta/L) = 3e^{-2X}. \tag{2.5}$$

(Note that Hegarty-Martinsson consider a different scaling regime, but their proof can be adapted without any modification to the result above.)

Since the variance in Proposition 1 scales like the square of the expectation, it means that when the starting value is $\mathcal{O}(1/L)$ there are $\mathcal{O}(L)$ open paths in the system. When the starting value is chosen randomly, there is a probability $\mathcal{O}(1/L)$ that it is in fact $\mathcal{O}(1/L)$ yielding $\mathcal{O}(L)$ open paths. On average we thus expect from these events alone $\mathcal{O}(1)$ open paths, with a variance $\mathcal{O}(L)$, as in (2.3).

This observation can be made more precise:

Theorem 1. *On the tree, for a starting value $x = X/L$, the variable θ/L converges in law when $L \to \infty$ to e^{-X} multiplied by a standard exponential variable.*

Theorem 2. *On the hypercube, for a starting value $x = X/L$, the variable θ/L converges in law when $L \to \infty$ to e^{-X} multiplied by the product of two independent standard exponential variables.*

It will become apparent in the proofs that we get a product of two independent variables on the hypercube because, locally near both corners $(0, 0, \dots, 0)$ and $(1, 1, \dots, 1)$, the hypercube graph looks roughly like the tree.

We conclude with the following remark: when the starting value x is picked randomly, even if the expectation and the variance of θ is dominated by values of $x = \mathcal{O}(1/L)$, the probability that there exists at least one open path is dominated by starting values $x = \ln L/L + \mathcal{O}(1/L)$. This was made clear on the hypercube in [2] and we here state the tree counterpart.

Theorem 3. *On the tree, when the starting value is $x = (\ln L + X)/L$*

$$\lim_{L\to\infty} \mathbb{E}^{\frac{\ln L+X}{L}}(\theta) = e^{-X}, \qquad \lim_{L\to\infty} \mathrm{Var}^{\frac{\ln L+X}{L}}(\theta) = e^{-2X} + e^{-X}. \tag{2.6}$$

When the starting value x is chosen at random uniformly in $[0, 1]$, the probability to have no open path goes to 1 as $L \to \infty$ and

$$\mathbb{P}^*(\theta \geq 1) \sim \frac{\ln L}{L} \qquad \text{as } L \to \infty. \tag{2.7}$$

The rest of the paper is organized as follows: we start by proving Theorem 3 in Section 3 as it is the simplest, and then we prove Theorem 1 in Section 4. The proofs rely on Proposition 1 which is itself proven in Section 5 for the tree. In Section 6, we introduce the notion of Poisson cascade, which allows to give a probabilistic interpretation of one

of the main object introduced in our proofs. Finally, Theorem 2 (on the hypercube) is proven in Section 7.

3. Proof of Theorem 3

In (2.6), the result on the expectation is trivial from (2.1), and the result on the variance is obtained in Section 5. In this section, we prove (2.7).

Let us start with the upper bound. Markov's inequality with (2.1) leads to

$$\mathbb{P}^x(\theta \geq 1) \leq \min\big[1, L(1-x)^{L-1}\big]. \tag{3.1}$$

We split the integral $\mathbb{P}^*(\theta \geq 1) = \int_0^1 \mathbb{P}^x(\theta \geq 1)\,\mathrm{d}x$ at $x_0 = 1 - \exp[-(\ln L)/(L-1)]$ since that is the point such that $L(1-x_0)^{L-1} = 1$. We end up with

$$\mathbb{P}^*(\theta \geq 1) \leq 1 - \exp\Big[-\frac{\ln L}{L-1}\Big] + \exp\Big[-\frac{L}{L-1}\ln L\Big] = \frac{\ln L}{L} + \mathcal{O}\Big(\frac{1}{L}\Big). \tag{3.2}$$

We now turn to the lower bound. Let $L \mapsto f(L)$ be a function diverging more slowly than $\ln L$:

$$\lim_{L\to\infty} f(L) = \infty, \qquad 0 \leq f(L) \leq \ln L, \qquad \lim_{L\to\infty} \frac{f(L)}{\ln L} = 0. \tag{3.3}$$

It is sufficient to show that

$$\lim_{L\to\infty} \mathbb{P}^{\frac{\ln L - f(L)}{L}}(\theta \geq 1) = 1, \tag{3.4}$$

because

$$\mathbb{P}^*(\theta \geq 1) \geq \int_0^{\frac{\ln L - f(L)}{L}} \mathbb{P}^x(\theta \geq 1)\,\mathrm{d}x \geq \frac{\ln L - f(L)}{L}\mathbb{P}^{\frac{\ln L - f(L)}{L}}(\theta \geq 1), \tag{3.5}$$

where we used that $x \mapsto \mathbb{P}^x(\theta \geq 1)$ is a non-increasing function. Taking L large in (3.5) assuming (3.4) gives the lower bound.

It now remains to show (3.4). We consider a tree started from $x = [\ln L - f(L)]/L$, and call m the number of nodes at first level with a value between x and $(\ln L)/L$. This number m is a binomial of parameters L and $f(L)/L$. Conditionally on m, the probability to have no open path in the tree is smaller than the probability to have no open path through these m specific nodes. Thus by summing over all possible values of m we get:

$$\mathbb{P}^{\frac{\ln L - f(L)}{L}}(\theta = 0) \leq \sum_{m=0}^{L} \binom{L}{m} \Big(\frac{f(L)}{L}\Big)^m \Big(1 - \frac{f(L)}{L}\Big)^{L-m} \Big[\mathbb{P}^{\frac{\ln L}{L};L-1}(\theta = 0)\Big]^m, \tag{3.6}$$

where we used $\mathbb{P}^x(\theta = 0) \leq \mathbb{P}^{(\ln L)/L}(\theta = 0)$ for $x \leq (\ln L)/L$. Note the obvious extension to the notation to mark that the probability on the right hand side is for a tree of size $L-1$ and not L as on the left hand side. Summing (3.6), one gets

$$\mathbb{P}^{\frac{\ln L - f(L)}{L}}(\theta = 0) \leq \Big[1 - \frac{f(L)}{L}\Big(1 - \mathbb{P}^{\frac{\ln L}{L};L-1}(\theta = 0)\Big)\Big]^L. \tag{3.7}$$

But from Cauchy-Schwarz (applied to θ and $\mathbb{1}(\theta \geq 1)$) and (2.6), which is proved in Section 5.3, one has

$$\mathbb{P}^{\frac{\ln L}{L}}(\theta \geq 1) \geq \frac{\mathbb{E}^{\frac{\ln L}{L}}(\theta)^2}{\mathbb{E}^{\frac{\ln L}{L}}(\theta^2)} \xrightarrow[L\to\infty]{} \frac{1}{3}, \tag{3.8}$$

so that, for L large enough

$$\mathbb{P}^{\frac{\ln L}{L};L-1}(\theta \geq 1) \geq \mathbb{P}^{\frac{\ln(L-1)}{L-1};L-1}(\theta \geq 1) \geq 0.33, \tag{3.9}$$

and thus, for L large enough

$$\mathbb{P}^{\frac{\ln L - f(L)}{L}}(\theta = 0) \leq \left[1 - \frac{f(L)}{L}\mathbb{P}^{\frac{\ln L}{L};L-1}(\theta \geq 1)\right]^L \leq \left[1 - 0.33\frac{f(L)}{L}\right]^L \tag{3.10}$$

which goes to zero as $L \to \infty$, as required.

4. Proof of Theorem 1

In this section, we consider the case of the tree with a starting value x which scales as $x = X/L$, $X \geq 0$ being a fixed number. The natural starting point for a proof would be to introduce G, the generating function of θ:

$$G(\lambda, x, L) = \mathbb{E}^x\left(e^{-\lambda\theta}\right), \tag{4.1}$$

with parameter $\lambda \geq 0$, for which it is very easy to show from the tree geometry (each of the L nodes at first level are the root of an independent tree of size $L-1$) that:

$$G(\lambda, x, 1) = e^{-\lambda}, \qquad G(\lambda, x, L) = \left[x + \int_x^1 \mathrm{d}y\, G(\lambda, y, L-1)\right]^L \quad \text{for } L > 1. \tag{4.2}$$

However, extracting the limiting distribution directly from (4.2) seems difficult because the number of levels and the size of each level increase together and because the fixed point equation does not give the λ dependence of the result. We shall rather use an idea which proved to be very generic and powerful in branching processes: the value of a random variable is decided during the early stages of a branching process; at later stages the law of large numbers kicks in (see for instance [5]).

Assume that all the information at the first k levels of the tree is known, and call θ_k the expected number of paths given that information:

$$\theta_k = \mathbb{E}(\theta|\mathcal{F}_k), \tag{4.3}$$

where $\mathcal{F}_k$ is the available information up to level k. For instance, consider the tree of Figure 1 up to level $k=2$. There are three paths still open with end values (at level 2) given by 0.22, 0.66 and 0.95. Therefore from (2.1), $\theta_2 = 2(1-0.22)+2(1-0.66)+2(1-0.95) = 2.34$. Similarly, $\theta_1 = 3(1-0.59)^2+3(1-0.90)^2+3(1-0.01)^2+3(1-0.83)^2 = 3.56$ and $\theta_3 = \theta_4 = \theta = 2$. A general expression of θ_k for $k < L$ is

$$\theta_k = \sum_{|\sigma|=k} \mathbb{1}_{\{\sigma \text{ open}\}}(L-k)(1-x_\sigma)^{L-k-1}, \tag{4.4}$$

where we sum over all nodes σ at level $|\sigma| = k$ in the tree, x_σ is the value of the node σ and the event $\{\sigma \text{ open}\}$ is the $\mathcal{F}_{|\sigma|}$-measurable event that the path from the root to node σ is open.

Heuristically, when k is small, there are few paths open up to level k and the variance of θ given $\mathcal{F}_k$ is large: θ has no reason to be close to θ_k. When k is large, however, there are many paths open up to level k which all contribute to the value of θ. The law of large numbers leads (on the good scale) to a small variance of θ given $\mathcal{F}_k$ and θ_k becomes a good approximation of θ. The advantage of this approach is that one can

take the $L \to \infty$ limit for a fixed k (keeping the depth of the tree constant) and then take the $k \to \infty$ limit.

Our proof consists then in two steps:

- first we show that

$$\lim_{L\to\infty} \mathbb{P}^{\frac{X}{L}} \left(\frac{\theta}{L} \le z \right) = \lim_{k\to\infty} \lim_{L\to\infty} \mathbb{P}^{\frac{X}{L}} \left(\frac{\theta_k}{L} \le z \right), \tag{4.5}$$

  which means that θ/L for a starting point X/L and L large has the same distribution as θ_k/L with the same starting point for large L and then for large k;
- then we make use of a generating function similar to (4.1) to show that the distribution of θ_k/L after taking the limits is given by an exponential law.

4.1. **Proof of** (4.5)**.** Pick $\delta > 0$. Observe that

$$\begin{aligned} \mathbb{P}\Big(\frac{\theta}{L} \le z \,\Big|\, \mathcal{F}_k\Big) &\le \mathbb{1}\Big(\frac{\theta_k}{L} \le z+\delta \,\Big|\, \mathcal{F}_k\Big) + \mathbb{P}\Big(\frac{|\theta-\theta_k|}{L} \ge \delta \,\Big|\, \mathcal{F}_k\Big), \\ \mathbb{P}\Big(\frac{\theta}{L} \le z \,\Big|\, \mathcal{F}_k\Big) &\ge \mathbb{1}\Big(\frac{\theta_k}{L} \le z-\delta \,\Big|\, \mathcal{F}_k\Big) - \mathbb{P}\Big(\frac{|\theta-\theta_k|}{L} \ge \delta \,\Big|\, \mathcal{F}_k\Big). \end{aligned} \tag{4.6}$$

The first inequality follows from the simple remark that it is obviously true when $\theta_k/L \le z+\delta$ and that, when $\theta_k/L > z+\delta$, it is necessary to have $(\theta_k-\theta)/L \ge \delta$ to get $\theta/L \le z$. Similarly, the lower bound in the next line is trivial when $\theta_k/L > z-\delta$ and it is sufficient to have $(\theta-\theta_k)/L < \delta$ when $\theta_k/L \le z-\delta$.

As θ_k is the expectation of θ given $\mathcal{F}_k$, one has from Chebyshev's inequality:

$$\mathbb{P}\Big(\frac{|\theta-\theta_k|}{L} \ge \delta \,\Big|\, \mathcal{F}_k\Big) \le \frac{\mathrm{Var}(\theta|\mathcal{F}_k)}{L^2\delta^2}. \tag{4.7}$$

We use (4.7) into (4.6) and then take the expectation (over $\mathcal{F}_k$):

$$\mathbb{P}^{\frac{X}{L}}\Big(\frac{\theta_k}{L} \le z-\delta\Big) - \frac{\mathbb{E}^{\frac{X}{L}}\big[\mathrm{Var}(\theta|\mathcal{F}_k)\big]}{L^2\delta^2} \le \mathbb{P}^{\frac{X}{L}}\Big(\frac{\theta}{L} \le z\Big) \le \mathbb{P}^{\frac{X}{L}}\Big(\frac{\theta_k}{L} \le z+\delta\Big) + \frac{\mathbb{E}^{\frac{X}{L}}\big[\mathrm{Var}(\theta|\mathcal{F}_k)\big]}{L^2\delta^2}.$$

Therefore to show (4.5), it is sufficient to have

$$\lim_{k\to\infty} \limsup_{L\to\infty} \frac{1}{L^2} \mathbb{E}^{\frac{X}{L}}\big[\mathrm{Var}(\theta|\mathcal{F}_k)\big] = 0, \tag{4.8}$$

as well as the existence and continuity of the right hand-side limit of (4.5).

In Section 5.5, we will show by direct analysis of the second moment that

$$\lim_{L\to\infty} \frac{1}{L^2} \mathbb{E}^{\frac{X}{L}}\big[\mathrm{Var}(\theta|\mathcal{F}_k)\big] = \frac{e^{-2X}}{2^k}, \tag{4.9}$$

which yields (4.8). We will now compute the distribution of θ_k/L in the double limit $L \to \infty$ and $k \to \infty$ and, as the result will be a continuous function of z, this will complete the proof.

4.2. **Distribution of θ_k.** Similarly to (4.1), we define $G_k(\lambda, x, L)$ the generating function of θ_k for a tree of size L and a value x at the root:

$$G_k(\lambda, x, L) = \mathbb{E}^x\left(e^{-\lambda\theta_k}\right). \tag{4.10}$$

As $\theta_0 = \mathbb{E}^x(\theta) = L(1-x)^{L-1}$ one has

$$G_0(\lambda, x, L) = \exp\left[-\lambda L(1-x)^{L-1}\right], \tag{4.11}$$

and the recursion relation

$$G_k(\lambda, x, L) = \left[x + \int_x^1 \mathrm{d}y\, G_{k-1}(\lambda, y, L-1)\right]^L = \left[1 - \int_x^1 \mathrm{d}y\left(1 - G_{k-1}(\lambda, y, L-1)\right)\right]^L, \tag{4.12}$$

to be compared to (4.2).

This relation is obtained by decomposing on what happen at the first splitting. For a node σ connected to the root let $\theta_k(\sigma)$ be the conditional expectation given $\mathcal{F}_k$ of the number of open paths going through σ. The $\{\theta_k(\sigma)\}_{|\sigma|=1}$ is a collection of L independent $\mathcal{F}_k$-measurable independent variables, hence:

$$G_k(\lambda, x, L) = \left[\mathbb{E}^x\big(e^{-\lambda\theta_k(\sigma)}\big)\right]^L, \tag{4.13}$$

where σ is a given node in the first generation.

Let us evaluate $\mathbb{E}^x\big(e^{-\lambda\theta_k(\sigma)}\big)$. If $x_\sigma < x$ then $\theta_k(\sigma) = 0$ and since this event has probability x it contribute $xe^{-\lambda 0} = x$ to the expectation. With a probability $\mathrm{d}y$ for $y \in [x, 1]$ the value at the node is $y > x$ and some paths might go through that node. The subtree rooted at σ is like the initial tree but of dimension $L-1$ and we want to evaluate the average number of paths in that subtree given the information after $k-1$ steps, hence the term in the integral of (4.12).

The strategy is to take the $L \to \infty$ limit at fixed k in (4.11) and (4.12) after a proper rescaling, then to let $k \to \infty$. We only consider $\lambda \geq 0$; it is sufficient to characterize the distribution, and it simplifies the arguments below.

Step 1. We first show that the following limit exists (for $\mu \geq 0$):

$$\forall a, b, \qquad G_k\left(\frac{\mu}{L+a}, \frac{X}{L+b}, L\right) \xrightarrow[L\to\infty]{} \widetilde{G}_k(\mu, X), \tag{4.14}$$

and that the limit satisfies

$$\widetilde{G}_k(\mu, X) = \exp\left[-\int_X^\infty [1 - \widetilde{G}_{k-1}(\mu, Y)]\mathrm{d}Y\right], \qquad \widetilde{G}_0(\mu, X) = \exp\left[-\mu e^{-X}\right]. \tag{4.15}$$

From (4.11), it is obvious that (4.14) holds for $k = 0$ with the limit given in (4.15). Choosing $k > 0$, we assume that (4.14) holds for G_{k-1}. Then, after a change of variable in (4.12),

$$G_k\left(\frac{\mu}{L+a}, \frac{X}{L+b}, L\right) = \left[1 - \frac{1}{L+b}\int_X^{L+b} \mathrm{d}Y\left(1 - G_{k-1}\Big(\frac{\mu}{L+a}, \frac{Y}{L+b}, L-1\Big)\right)\right]^L.$$

The G_{k-1} on the right hand side has a $L \to \infty$ limit. From its definition (4.10), one has

$$1 \geq G_k(\lambda, x, L) \geq 1 - \lambda\mathbb{E}^x(\theta_k) = 1 - \lambda\mathbb{E}^x(\theta) = 1 - \lambda L(1-x)^{L-1}. \tag{4.16}$$

Then, assuming $\mu \geq 0$, for all a and b, one has for L large enough (depending on a and b):

$$1 \geq G_{k-1}\Big(\frac{\mu}{L+a}, \frac{Y}{L+b}, L-1\Big) \geq 1 - 2\mu e^{-Y/2}. \tag{4.17}$$

Thus, from the dominated convergence theorem, we have

$$\int_X^{L+b} \mathrm{d}Y \left(1 - G_{k-1}\Big(\frac{\mu}{L+a}, \frac{Y}{L+b}, L-1\Big)\right) \xrightarrow[L\to\infty]{} \int_X^\infty \mathrm{d}Y \big(1 - \widetilde{G}_{k-1}(\mu, Y)\big), \tag{4.18}$$

and thus (4.14) holds for G_k with the relation (4.15).

Step 2. The fact that (4.14) holds means that when starting with $x = X/L$, the random variable θ_k/L has a well defined limit as L goes to infinity, and that the generating function of that limit is $\widetilde{G}_k$. We now use the recurrence (4.15) to take the $k \to \infty$ limit which will show that $\lim_{L\to\infty} \theta_k/L$ converges (when $k \to \infty$) to an exponential variable. This task is greatly simplified by noticing (by a simple recurrence) that one can write $\widetilde{G}_k$ as a function of one variable only:

$$\widetilde{G}_k(\mu, X) = F_k(\mu e^{-X}) \tag{4.19}$$

with

$$F_k(z) = \exp\left[-\int_0^z \frac{1 - F_{k-1}(z')}{z'} \mathrm{d}z'\right], \qquad F_0(z) = e^{-z}. \tag{4.20}$$

We shall show that the solution to (4.20) satisfies

$$F_k(z) \xrightarrow[k\to\infty]{} \frac{1}{1+z} \qquad \text{for } z > -1, \tag{4.21}$$

which implies that $\lim_{L\to\infty} \theta_k/L$ converges weakly when $k \to \infty$ to an exponential distribution of expectation e^{-X}. Note that we only need to consider $z \geq 0$ and, in fact, we proved (4.20) only for $z \geq 0$, but (4.21) holds for the solution to (4.20) for $z \in (-1, \infty)$.

Defining $\delta_k(z)$ for $z > -1$ and $z \neq 0$ by

$$F_k(z) = \frac{1}{1+z} - \frac{z^2}{(1+z)^3}\,\frac{\delta_k(z)}{2^k}, \tag{4.22}$$

it is easy to see that there exists a constant M such that for all k and all $z > -1$

$$0 \leq \delta_k(z) \leq M. \tag{4.23}$$

Indeed, for $k = 0$,

$$\delta_0(z) = \frac{(1+z)^3}{z^2}\left(\frac{1}{1+z} - e^{-z}\right), \tag{4.24}$$

$\delta_0(z) \geq 0$ for $z > -1$ because $e^z \geq 1 + z$ by convexity. Furthermore, $\delta_0(z)$ can be defined by continuity at $z = 0$, has a limit in $z = +\infty$ and in $z = -1$ and reaches therefore a maximum M on $(-1, \infty)$, which initializes (4.23).

Assuming now (4.23) at order $k - 1$, one has

$$F_k(z) = \frac{1}{1+z} \exp\left[-\int_0^z \frac{z'}{(1+z')^3}\,\frac{\delta_{k-1}(z')}{2^{k-1}} \mathrm{d}z'\right], \tag{4.25}$$

leading to

$$\frac{1}{1+z} \geq F_k(z) \geq \frac{1}{1+z}\left[1 - \frac{M}{2^{k-1}}\int_0^z \frac{z'}{(1+z')^3}\,\mathrm{d}z'\right] = \frac{1}{1+z}\left[1 - \frac{M}{2^{k-1}}\,\frac{z^2}{2(1+z)^2}\right],$$

which gives (4.23) at order k. Hence the limit (4.21) holds. This completes the proof of Theorem 1.

5. Results on the second moment for the tree

The goal of this section is to prove the second moment results (2.3), (2.4), (2.6) and (4.9) which were used in the proofs of Theorems 1 and 3.

5.1. **Exact expression of the second moment.** The expectation of θ^2 is the sum, over all the pairs of paths, of the probability that both paths are open. There are $L!^2$ pairs of paths in the system. For a given pair, the probability that they are both open depends on the number $q \in \{0,1,2,\dots,L-2,L\}$ of bonds shared by the paths. (Nota: two paths may not have exactly $L-1$ bonds in common.) The following facts are clear:

- the number of pairs of paths which coincide all the way ($q = L$) is $L!$;
- the probability that "both" paths in such a pair are open is $(1-x)^{L-1}/(L-1)!$;
- the number of pairs of paths which coincide for $q = 0,1,\dots L-2$ steps and then branch is $L!(L-q-1)(L-q-1)!$; (Remark: $1\ 1!+2\ 2!+3\ 3!+\cdots+(L-1)(L-1)! = L!-1$, hence one recovers that the total number of pairs of path is $L!^2$.)
- the probability that both paths in such a pair are open is
$$\frac{(1-x)^{2L-q-2}}{(2L-q-2)!}\binom{2L-2q-2}{L-q-1}. \tag{5.1}$$
Indeed, excluding the starting and end points, there are $2L-q-2$ total different nodes in such a pair of paths. All these nodes must be larger than x, hence the $(1-x)^{2L-q-2}$ term. This is however not sufficient because the values on the nodes must be correctly ordered. Out of the $(2L-q-2)!$ possible ordering (see the denominator), the only good ones are those such that the q smallest terms are well ordered in the shared segment (only one choice), and the $2L-2q-2$ remaining terms are separated into two well ordered blocks of $L-q-1$ terms, one for each path; the only freedom is to choose which terms go to which path, hence the binomial coefficient.

This leads to

$$\mathbb{E}^x(\theta^2) = \sum_{q=0}^{L-2} a(L,q)(1-x)^{2L-q-2} + L(1-x)^{L-1}, \tag{5.2}$$

where

$$a(L,q) = \frac{L!(2L-2q-2)!}{(L-q-2)!(2L-q-2)!}. \tag{5.3}$$

The isolated term in (5.2) corresponds to the pairs of identical paths and is equal to $\mathbb{E}^x(\theta)$.

5.2. **Estimates and bounds on the** $a(L,q)$**.** Expanding the factorials in $a(L,q)$, one gets

$$a(L,q) = \frac{L^2}{2^q}\,\frac{\Big(1-\frac{1}{L}\Big)\Big(1-\frac{2}{L}\Big)\cdots\Big(1-\frac{q+1}{L}\Big)}{\Big(1-\frac{q+2}{2L}\Big)\Big(1-\frac{q+3}{2L}\Big)\cdots\Big(1-\frac{2q+1}{2L}\Big)}. \tag{5.4}$$

From this expression, one gets the following equivalent when $L\to\infty$ and $q \ll \sqrt{L}$:

$$a(L,q) = \frac{L^2}{2^q}\big[1+\mathcal{O}(q^2/L)\big]. \tag{5.5}$$

For q close to L, one has the values $a(L,L-2)=2$, $a(L,L-3)=24/L$, $a(L,L-4)=360/[(L+1)(L+2)]$, etc.

We want to find a good upper bound for $a(L,q)$. We first show that $q\mapsto \ln a(L,q)$ is a convex function for $L\ge q+3$. Indeed

$$\ln a(L,q) - \ln a(L,q-1) = \ln\frac{(L-q-1)(2L-q-1)}{(2L-2q)(2L-2q-1)} \tag{5.6}$$

so that

$$\begin{aligned}&[\ln a(L,q)-\ln a(L,q-1)] - [\ln a(L,q-1) - \ln a(L,q-2)]\\ &\qquad = \ln\frac{(L-q-1)(2L-q-1)(2L-2q+2)(2L-2q+1)}{(2L-2q)(2L-2q-1)(L-q)(2L-q)}.\end{aligned} \tag{5.7}$$

Since the denominator is clearly positive as soon as $L\ge q+1$ we see that $\ln a(L,q)$ is convex if in (5.7) the numerator is bigger than the denominator. This condition leads to

$$(L-q)^2(2L-1) - (2L-q-1)(2L-2q+1) \ge 0, \tag{5.8}$$

which holds as soon as $L-q\ge 3$.

Assume $L\ge 12$ so that $a(L,L-3)\le 2$ and let

$$q_0(L) = \left\lceil \frac{\ln(L^2)}{\ln 2} + 1 \right\rceil. \tag{5.9}$$

From (5.4) one has

$$a(L,q) \le \frac{L^2}{2^q}\,\frac{1}{(1-\frac{2q+1}{2L})^q}, \tag{5.10}$$

Applying this to $q=q_0(L)$, one easily gets $a(L,q_0(L))\le 1$ for L large enough. (The term $L^2/2^{q_0}$ is smaller than $1/2$, and the parenthesis converges to 1.)

By using the convexity of $\ln a(L,q)$, one has

$$\ln a(L,q) \le \begin{cases} \ln a(L,0) + \frac{q}{q_0(L)}\Big[\ln a(L,q_0(l)) - \ln a(L,0)\Big] & \text{for } 0\le q\le q_0(L),\\ \ln 2 & \text{for } q_0(L)\le q\le L-2.\end{cases}$$

But $\ln a(L,q_0(L))\le 0$ so that

$$a(L,q) \le \begin{cases} a(L,0)\exp\Big[-\frac{\ln a(L,0)}{q_0(L)}q\Big] & \text{for } 0\le q\le q_0(L),\\ 2 & \text{for } q_0(L)\le q\le L-2.\end{cases} \tag{5.11}$$

Remark now that $a(L,0)=L(L-1)<L^2$ and $\ln a(L,0)/q_0(L)\to\ln 2$ as $L\to\infty$. This implies that, for L large enough, $\ln a(L,0)/q_0(L)>\ln 1.99$ and that

$$a(L,q)\le\begin{cases}L^2 1.99^{-q} & \text{for } 0\le q\le q_0(L),\\ 2 & \text{for } q_0(L)\le q\le L-2.\end{cases}\tag{5.12}$$

5.3. **Proofs of the limits** (2.3)**,** (2.4) **and** (2.6) **of** $\mathrm{Var}(\theta)$**.** The second moment (5.2) is written as a sum from $q=0$ to $q=L-2$. To prove the various limits we need, the strategy is always the same:

(1) Split the sum over q into two parts; one going from 0 to $q_0(L)$ and one going from $q_0(L)+1$ to $L-2$.
(2) In the first sum, replace $a(L,q)$ by its equivalent (5.5); this is justified with the dominated convergence theorem, using the bound (5.12).
(3) Show that the second sum do not contribute using the bound (5.12).

Proof of (2.3). Integrating (5.2) over x and using $\mathbb{E}^*(\theta)=1$, one gets

$$\frac{\mathrm{Var}^*(\theta)}{L}=\frac{\mathbb{E}^*(\theta^2)-\mathbb{E}^*(\theta)^2}{L}=\sum_{q=0}^{L-2}\frac{a(L,q)/L}{2L-q-1}=\sum_{q=0}^{q_0(L)}\frac{a(L,q)/L^2}{2-\frac{q+1}{L}}+\sum_{q=q_0(L)+1}^{L-2}\frac{a(L,q)/L}{2L-q-1}.$$

In the first sum, the running term is equivalent to $2^{-q}/2$ when L is large and is dominated by 1.99^{-q} for L large enough. Therefore, this first sum converges to 1. In the second sum, the running term is smaller than $2/L^2$, implying that the whole second sum is smaller than $2/L$ and thus vanishes in the large L limit.

Proof of (2.4). We divide (5.2) by L^2, replace x by X/L and split the sum:

$$\frac{\mathbb{E}^{\frac{X}{L}}(\theta^2)}{L^2}=\sum_{q=0}^{q_0(L)}\frac{a(L,q)}{L^2}\Big(1-\frac{X}{L}\Big)^{2L-q-2}+\sum_{q=q_0(L)+1}^{L-2}\frac{a(L,q)}{L^2}\Big(1-\frac{X}{L}\Big)^{2L-q-2}+\frac{(1-\frac{X}{L})^{L-1}}{L}.\tag{5.13}$$

The running term in the first sum is equivalent to $2^{-q}e^{-2X}$ and is dominated by 1.99^{-q}, therefore the first term converges to $2e^{-2X}$ as $L\to\infty$. The running term in the second sum is smaller than $2/L^2$ implying that the whole second sum is smaller than $2/L$ and thus vanishes in the large L limit. The isolated term goes also to zero. Therefore, the whole expression converges to $2e^{-2X}$ and one recovers the variance in (2.4) after subtracting $\mathbb{E}^{X/L}(\theta/L)^2$.

Proof of (2.6). We now take $x=(\ln L+X)/L$ and split again the sum in (5.2) into two parts:

$$\mathbb{E}^x(\theta^2)=\sum_{q=0}^{q_0(L)}\frac{a(L,q)}{L^2}\times L^2(1-x)^{2L-q-2}+\sum_{q=q_0(L)+1}^{L-2}a(L,q)(1-x)^{2L-q-2}+L(1-x)^{L-1}.\tag{5.14}$$

Using

$$\lim_{L\to\infty}L^2\Big(1-\frac{\ln L+X}{L}\Big)^{2L-q-2}=e^{-2X},\tag{5.15}$$

into (5.14), the running term in the first sum is equivalent to $2^{-q}e^{-2X}$ and is dominated by $1.99^{-q}(e^{-2X}+1)$ for L large enough (because $L^2(1-x)^{2L-q-2}\le L^2(1-x)^{2L-q_0(L)-2}$,

which becomes close to its limit when L gets large). Therefore, the first sum converges to $2e^{-2X}$. We write an upper bound of the second sum of (5.14) using $a(L,q) \le 2$ and then extending the sum to the interval $[0, L-1]$:

$$\sum_{q=q_0(L)+1}^{L-2} a(L,q)(1-x)^{2L-q-2} \le 2(1-x)^{2L-2}\frac{(1-x)^{-L}-1}{(1-x)^{-1}-1} \le 2\frac{(1-x)^{L-1}}{x} \sim 2\frac{e^{-X}}{\ln L}, \tag{5.16}$$

which goes to zero for L large. Finally, the last term in (5.14) converges to e^{-X}; putting things together, one finds $\mathbb{E}^{(\ln L+X)/L}(\theta^2) \to 2e^{-2X} + e^{-X}$. Removing the expectation squared, one recovers (2.6).

5.4. **Exact expression for** $\mathbb{E}^x[\mathrm{Var}(\theta|\mathcal{F}_k)]$**.** The number θ of paths given $\mathcal{F}_k$ is the sum over all the nodes at level k of the number of paths through that node. These variables are independent; therefore

$$\mathrm{Var}(\theta|\mathcal{F}_k) = \sum_{|\sigma|=k} \mathbb{1}_{\{\sigma \text{ open}\}} v(x_\sigma, L-k) \tag{5.17}$$

where $v(x,L)$ is the variance of θ for a tree of size L started at x

$$v(x,L) := \mathbb{E}^x(\theta^2) - \mathbb{E}^x(\theta)^2 = -L(1-x)^{2L-2} + \sum_{q=1}^{L-2} a(L,q)(1-x)^{2L-q-2} + L(1-x)^{L-1}. \tag{5.18}$$

Taking the expectation over $\mathcal{F}_k$, one gets

$$\mathbb{E}^x[\mathrm{Var}(\theta|\mathcal{F}_k)] = \frac{L!}{(L-k)!}\int_x^1 \mathrm{d}x_\sigma\, \frac{(x_\sigma - x)^{k-1}}{(k-1)!} v(x_\sigma, L-k), \tag{5.19}$$

where $L!/(L-k)!$ is the number of terms in the sum and where the fraction in the integral is the probability that σ is open given the value of $x_\sigma > x$.

Performing the integration term by term is easy, one finds, after simplification:

$$\mathbb{E}^x[\mathrm{Var}(\theta|\mathcal{F}_k)] = -\frac{a(L,k)}{L-k-1}(1-x)^{2L-k-2} + \sum_{q=k+1}^{L-2} a(L,q)(1-x)^{2L-q-2} + L(1-x)^{L-1}. \tag{5.20}$$

Note that apart from the first term, this is exactly the same as the full variance $v(x,L)$ except that the sum over q begins at $k+1$ instead of at 1.

5.5. **Proof of** (4.9)**.** We now divide (5.20) by L^2, set $x = X/L$ and consider L large. We only need an upper bound, but it is as easy to calculate the exact limit. As in Section 5.3, we split the sum into two parts; one where the index q runs from k to $q_0(L)$ and one from $q_0(L)+1$ to $L-2$. In the first part, using the dominated convergence theorem with the bound (5.12):

$$\lim_{L\to\infty} \sum_{q=k+1}^{q_0(L)} \frac{a(L,q)}{L^2}\Big(1-\frac{X}{L}\Big)^{2L-q-2} = \sum_{q=k+1}^{\infty} \frac{1}{2^q} e^{-2X} = \frac{1}{2^k}e^{-2X}. \tag{5.21}$$

Also using the bound (5.12), the second part of the sum goes to zero:

$$\frac{1}{L^2}\sum_{q=q_0(L)+1}^{L-2} a(L,q)\Big(1-\frac{X}{L}\Big)^{2L-q-2} \le \frac{1}{L^2}\times L\times 2. \tag{5.22}$$

It is very easy to check that in (5.20) the two isolated terms (divided by L^2, of course) go also to zero, so that one finally obtains (4.9).

6. A relation with Poisson cascades

Our model is closely related to cascades of Poisson processes. In fact, the arguments we used in Section 3 can be presented in terms of Poisson cascades. Let us make a brief description.

We recall the sequence of functions F_k, $k\ge 0$, defined in (4.20):

$$F_k(z) = \exp\left[-\int_0^z \frac{1-F_{k-1}(z')}{z'}\mathrm{d}z'\right], \qquad F_0(z)=e^{-z}. \tag{6.1}$$

It is clear that $F_0(z)$ is the Laplace transform of the Dirac mass at 1, and that F_1 is the Laplace transform of $\sum_{j=1}^\infty X_j$, where $(X_j,\, j\ge 1)$ is a Poisson process on (0, 1] with intensity $\mathbb{1}_{(0,\,1]}(x)\,\frac{\mathrm{d}x}{x}$.

We now define a cascade of Poisson processes. At generation $k=0$, there is only one particle at position 1. At generation $k=1$, this particle is replaced by the atoms $(X_j^{(1)},\, j\ge 1)$ of a Poisson process on (0, 1] with intensity $\mathbb{1}_{(0,\,1]}(x)\,\frac{\mathrm{d}x}{x}$. At generation $k=2$, for each j, the particle at position $X_j^{(1)}$ is replaced by $(X_j^{(1)}X_{j,\ell}^{(2)},\, \ell\ge 1)$, where $(X_{j,\ell}^{(2)},\, \ell\ge 1)$ is another Poisson process with intensity $\mathbb{1}_{(0,\,1]}(x)\,\frac{\mathrm{d}x}{x}$ (all the Poisson processes are assumed to be independent). Iterating the procedure results in a cascade of Poisson processes. We readily check, by induction in k, that F_k is the Laplace transform of Y_k, the sum of the positions at the k-th generation of the Poisson cascade.

What was proved in Section 3 can be stated in terms of the cascade of Poisson processes. Recall from (4.3) that $\theta_k = \mathbb{E}(\theta|\mathcal{F}_k)$.

Theorem 4. (i) *For any $k\ge 0$ and for $x = X/L$, $\frac{\theta_k}{L}$ converges weakly, when $L\to\infty$, to $e^{-X}Y_k$.*
(ii) *When $k\to\infty$, Y_k converges weakly to the standard exponential law.*

7. Proof of Theorem 2

In this section, we adapt the methods used in Section 4 to obtain the distribution of the number of open paths on the hypercube when L goes to infinity.

In the large L limit, both the width (the number of possible moves at each step) and the depth (the number of steps) on the hypercube go to infinity, which makes studying the limit difficult. We worked around that problem on the tree by introducing θ_k, the expected number of paths given the information $\mathcal{F}_k$ after k steps, and by sending first L (now representing only the width of the tree) and then k (the depth) to infinity.

We use the same trick on the hypercube, but with a twist: the hypercube is symmetrical when exchanging the starting and end points, and there is no reason to privilege

one or the other. Therefore, we call θ_k the expected number of paths in the hypercube given the information $\mathcal{F}_k$ at the first k levels *from both extremities of the hypercube.*

To write an expression for θ_k similar to (4.4), we introduce the following notations:

- The $\binom{L}{k}$ nodes k steps away from the starting point are indexed by σ and, as usual, their values are written x_σ.
- Similarly, τ indexes the $\binom{L}{k}$ nodes k steps away from the end point and we note their values $1 - y_\tau$.
- $n_\sigma \in \{0, 1, \dots, k!\}$ is the number of open paths from the starting point to node σ. (Contrary to the tree, there are several paths leading to each node σ.)
- Similarly, m_τ is the number of open paths from node τ to the end point.
- $\mathbb{1}(\sigma \leftrightarrow \tau)$ indicates whether there is at least one directed path (open or not) from node σ to node τ.

Then,

$$\theta_k = \sum_{|\sigma|=k} \sum_{|\tau|=L-k} n_\sigma m_\tau \mathbb{1}(\sigma \leftrightarrow \tau)(L-2k)(1-y_\tau - x_\sigma)^{L-2k-1}\mathbb{1}(x_\sigma + y_\tau \le 1), \tag{7.1}$$

where $\mathbb{1}(\sigma \leftrightarrow \tau)(L-2k)(1-y_\tau - x_\sigma)^{L-2k-1}\mathbb{1}(x_\sigma + y_\tau < 1)$ is the expected number of open paths from σ to τ given the values x_σ and y_τ.

Our proof can be decomposed into three steps:

- First, we show that, as in the hypercube, the distribution of θ/L as $L \to \infty$ is the same as the distribution of θ_k/L as $L \to \infty$ and then $k \to \infty$.
- Then, we show that the double sum in (7.1) can be modified (without changing the limit, of course) into a product of two sums. This means that asymptotically θ_k can be written as a contribution from the k first levels (the sum on σ) times an independent contribution from the k last levels (the sum on τ).
- Finally, we show that each of these two contributions is asymptotically identical in distribution to what we computed on the tree.

7.1. **First step: θ_k and θ have asymptotically the same distribution.** We show in this section that, when the starting point scales with L as $x = X/L$ for X fixed,

$$\lim_{L\to\infty} \frac{\theta_k}{L} \xrightarrow[k\to\infty]{\text{weakly}} \lim_{L\to\infty} \frac{\theta}{L}. \tag{7.2}$$

Following the same argument as on the tree, it is sufficient to show that

$$\lim_{k\to\infty} \limsup_{L\to\infty} \frac{1}{L^2} \mathbb{E}^{\frac{X}{L}}\big[\operatorname{Var}(\theta|\mathcal{F}_k)\big] = 0. \tag{7.3}$$

First remark that

$$\mathbb{E}^x\big[\operatorname{Var}(\theta|\mathcal{F}_k)\big] = \mathbb{E}^x[\theta^2] - \mathbb{E}^x[\theta_k^2] \tag{7.4}$$

where we used $\theta_k = \mathbb{E}[\theta|\mathcal{F}_k]$.

Second moments as in (7.4) can be written as sums over pairs of paths. For a given path α, we call x_i^α the value on the node at step i on path α ($0 \le i \le L$, with $x_0^\alpha = x$ and $x_L^\alpha = 1$) and $\xi_{i,j}^\alpha$ the indicator function that path α is open from steps i to j:

$$\xi_{i,j}^\alpha = \mathbb{1}(x_i^\alpha \le x_{i+1}^\alpha \le x_{i+2}^\alpha \le \dots \le x_j^\alpha). \tag{7.5}$$

Clearly,

$$\theta = \sum_\alpha \xi^\alpha_{0,L}, \qquad \theta_k = \sum_\alpha \mathbb{E}\big[\xi^\alpha_{0,L}|\mathcal{F}_k\big]. \tag{7.6}$$

We now have the following expression for the second moment:

$$\mathbb{E}^x\big[\theta^2\big] = \sum_{\alpha,\beta} \mathbb{E}^x\big[\xi^\alpha_{0,L}\xi^\beta_{0,L}\big] = L! \sum_\alpha \mathbb{E}^x\big[\xi^\alpha_{0,L}\xi^0_{0,L}\big] \tag{7.7}$$

where, by symmetry, we chose one particular arbitrary fixed path which bears the index 0. Similarly,

$$\mathbb{E}^x[\theta_k^2] = L! \sum_\alpha \mathbb{E}^x\Big[\mathbb{E}\big[\xi^\alpha_{0,L}|\mathcal{F}_k\big]\mathbb{E}\big[\xi^0_{0,L}|\mathcal{F}_k\big]\Big]. \tag{7.8}$$

We write now $\xi^\alpha_{0,L} = \xi^\alpha_{0,k}\xi^\alpha_{k,L-k}\xi^\alpha_{L-k,L}$. The first and last terms are $\mathcal{F}_k$-measurable, hence

$$\mathbb{E}^x[\theta_k^2] = L! \sum_\alpha \mathbb{E}^x\Big[\xi^\alpha_{0,k}\xi^0_{0,k}\,\mathbb{E}\big[\xi^\alpha_{k,L-k}|\mathcal{F}_k\big]\mathbb{E}\big[\xi^0_{k,L-k}|\mathcal{F}_k\big]\,\xi^\alpha_{L-k,L}\xi^0_{L-k,L}\Big]. \tag{7.9}$$

We make the same decomposition on $\xi^\alpha_{0,L}$ in (7.7). Writing $\mathbb{E}^x[\cdot] = \mathbb{E}^x[\mathbb{E}[\cdot|\mathcal{F}_k]]$ and pushing out of the inner expectation the $\mathcal{F}_k$-measurable terms, one gets

$$\mathbb{E}^x[\theta^2] = L! \sum_\alpha \mathbb{E}^x\Big[\xi^\alpha_{0,k}\xi^0_{0,k}\,\mathbb{E}\big[\xi^\alpha_{k,L-k}\xi^0_{k,L-k}|\mathcal{F}_k\big]\,\xi^\alpha_{L-k,L}\xi^0_{L-k,L}\Big]. \tag{7.10}$$

Using (7.4),

$$\mathbb{E}^x\big[\operatorname{Var}(\theta|\mathcal{F}_k)\big] = L!\times$$
$$\sum_\alpha \mathbb{E}^x\bigg[\xi^\alpha_{0,k}\xi^0_{0,k}\Big(\mathbb{E}\big[\xi^\alpha_{k,L-k}\xi^0_{k,L-k}|\mathcal{F}_k\big] - \mathbb{E}\big[\xi^\alpha_{k,L-k}|\mathcal{F}_k\big]\mathbb{E}\big[\xi^0_{k,L-k}|\mathcal{F}_k\big]\Big)\xi^\alpha_{L-k,L}\xi^0_{L-k,L}\bigg]. \tag{7.11}$$

For a given path α, the central term (in parenthesis) in the last expression is a kind of covariance. Clearly, if the paths α and 0 do not meet in the interval $\{k,\dots,L-k\}$, the variables $\xi^\alpha_{k,L-k}$ and $\xi^0_{k,L-k}$ are independent and the covariance is zero. Therefore, *we can restrict the sum over* α *in* (7.11) *to the paths which cross at least once the path* 0 *in the interval* $\{k,\dots,L-k\}$. With this modified sum, we can now find an upper bound to (7.4). Dropping all the negative terms and undoing the decomposition of $\xi^\alpha_{0,L}$ into three parts, we get

$$\mathbb{E}^x\big[\operatorname{Var}(\theta|\mathcal{F}_k)\big] \le L! {\sum_\alpha}' \mathbb{E}^x\big[\xi^\alpha_{0,L}\xi^0_{0,L}\big], \tag{7.12}$$

where the prime on the sum indicates that α runs only over all the paths that meet path 0 at least once in $\{k,\dots,L-k\}$.

We now bound (7.12). Let $I_{p,q}$ be the set of all the paths such that

- the $p+1$ first nodes (including the origin) are the same as for path 0 (in other words, the first p steps are the same as in path 0),
- the next $L-p-q-1$ nodes are different from those of path 0,
- the next $q+1$ nodes (thus including the end point) are the same as for path 0.

By construction, for $p<k$ and $q<k$, a path in $I_{p,q}$ do not meet path 0 in $\{k,\dots,L-k\}$. Therefore

$$\mathbb{E}^x\big[\operatorname{Var}(\theta|\mathcal{F}_k)\big] \le L! \sum_\alpha \mathbb{E}^x\Big[\xi^\alpha_{0,L}\xi^0_{0,L}\Big] - L! \sum_{p=0}^{k-1}\sum_{q=0}^{k-1}\sum_{\alpha\in I_{p,q}} \mathbb{E}^x\Big[\xi^\alpha_{0,L}\xi^0_{0,L}\Big]. \tag{7.13}$$

Notice that the first sum is not primed; it runs over all the $L!$ possible paths α. The inequality holds because in (7.12) we were summing over all the paths except *all* of those not crossing path 0 in $\{k, L-k\}$, while in (7.13) we sum over all the paths except *some* of those not crossing path 0 in $\{k, L-k\}$.

The first term in (7.13) is simply $\mathbb{E}^x[\theta^2]$, see (7.7). From [2] and (2.5) it has the following large L limit

$$\lim_{L\to\infty} \frac{1}{L^2}\mathbb{E}^{\frac{X}{L}}[\theta^2] = \lim_{L\to\infty} \frac{1}{L^2} L! \sum_{\alpha} \mathbb{E}^{\frac{X}{L}}\left[\xi^{\alpha}_{0,L}\xi^{0}_{0,L}\right] = 4e^{-2X}. \tag{7.14}$$

We now focus on the second term. For a path α in $I_{p,q}$, a direct calculation shows that

$$\mathbb{E}^x\left[\xi^{\alpha}_{0,L}\xi^{0}_{0,L}\right] = \frac{(1-x)^{2L-p-q-2}}{(2L-p-q-2)!}\binom{2L-2p-2q-2}{L-p-q-1}. \tag{7.15}$$

Indeed, excluding the starting and ending points, there are $2(L-1)-p-q$ total different nodes in the paths α and 0. All these nodes must be larger than x, hence the $(1-x)^{2L-p-q-2}$ term. This is however not sufficient because the values on the nodes must be correctly ordered. Out of the $(2L-p-q-2)!$ possible orderings (see the denominator), the only good ones are those such that the p smallest terms be well ordered in the first shared segment (only one choice), the q largest terms be well ordered in the second shared segment (only one choice), and the $2L-2p-2q-2$ remaining terms be separated into two well ordered blocks of $L-p-q-1$ terms, one for each path; the only freedom is to choose which terms go to path α and which to path 0, hence the binomial coefficient.

The number of terms in $I_{p,q}$ is $B(L-p-q)$, where $B(n)$ is the number of permutations of n elements such that for any m in $\{1,\dots,n-1\}$ the image of $\{1,\dots,m\}$ through the permutation is not $\{1,\dots,m\}$ (see [9] ; $B(1)=1$, $B(2)=1$, $B(3)=3$, $B(4)=13$, $B(5)=71$, ...). Hegarty Martinsson [2] call this $T(n,1)$ and show (Proposition 2.6) that $B(n)\sim n!$. Then

$$L! \sum_{\alpha\in I_{p,q}} \mathbb{E}^x\left[\xi^{\alpha}_{0,L}\xi^{0}_{0,L}\right] =$$
$$\frac{L!}{(L-p-q-1)!} \times \frac{(2L-2p-2q-2)!}{(2L-p-q-2)!} \times \frac{B(L-p-q)}{(L-p-q-1)!} \times (1-x)^{2L-p-q-2}.$$

Take $x = X/L$ and L large with p and q fixed. The terms on the right hand side are respectively equivalent to L^{p+q+1}, $(2L)^{-p-q}$, L and e^{-2X}, so that

$$\lim_{L\to\infty} \frac{1}{L^2} L! \sum_{\alpha\in I_{p,q}} \mathbb{E}^{\frac{X}{L}}\left[\xi^{\alpha}_{0,L}\xi^{0}_{0,L}\right] = \frac{e^{-2X}}{2^{p+q}}, \tag{7.16}$$

and

$$\lim_{L\to\infty} \frac{1}{L^2} L! \sum_{p=0}^{k-1}\sum_{q=0}^{k-1}\sum_{\alpha\in I_{p,q}} \mathbb{E}^{\frac{X}{L}}\left[\xi^{\alpha}_{0,L}\xi^{0}_{0,L}\right] = e^{-2X}4(1-2^{-k+1}+4^{-k}). \tag{7.17}$$

Using (7.14) and (7.17) in (7.13), we finally get

$$\limsup_{L\to\infty} \frac{1}{L^2}\mathbb{E}^{\frac{X}{L}}\left[\operatorname{Var}(\theta|\mathcal{F}_k)\right] \le \frac{8e^{-2X}}{2^k}, \tag{7.18}$$

from which one gets (7.3) and (7.2).

7.2. **Second step: separating the start and the end of the hypercube.** We go back to the expression θ_k given in (7.1):

$$\theta_k = \sum_{|\sigma|=k} \sum_{|\tau|=L-k} n_\sigma m_\tau \mathbb{1}(\sigma \leftrightarrow \tau)(L-2k)(1-y_\tau-x_\sigma)^{L-2k-1}\mathbb{1}(x_\sigma+y_\tau \le 1), \quad (7.19)$$

and we introduce the following slightly different quantity

$$\tilde\theta_k = \sum_{|\sigma|=k} \sum_{|\tau|=L-k} n_\sigma m_\tau L(1-y_\tau-x_\sigma+x_\sigma y_\tau)^{L-2k-1}. \quad (7.20)$$

(Compared to θ_k, this one has no $\mathbb{1}(\sigma \leftrightarrow \tau)$, no $\mathbb{1}(x_\sigma+y_\tau \le 1)$, a factor L instead of $L-2k$ and an extra $x_\sigma y_\tau$ in the power.) Clearly, $\theta_k \le \tilde\theta_k$. Furthermore, we know that $\mathbb{E}^x[\theta_k] = \mathbb{E}^x[\theta] = L(1-x)^{L-1}$ so that

$$\lim_{L\to\infty} \mathbb{E}^{\frac{X}{L}}\left[\frac{\theta_k}{L}\right] = e^{-X}. \quad (7.21)$$

Let us compute the same expectation for $\tilde\theta_k$. Using

$$\mathbb{E}^x(n_\sigma|x_\sigma) = k(x_\sigma-x)^{k-1}\mathbb{1}(x_\sigma \ge x), \qquad \mathbb{E}^x(m_\tau|y_\tau) = k(y_\tau)^{k-1}, \quad (7.22)$$

one gets

$$\begin{aligned}\mathbb{E}^x\left[\frac{\tilde\theta_k}{L}\right] &= \binom{L}{k}\binom{L}{k}\int_x^1 \mathrm{d}x_\sigma \int_0^1 \mathrm{d}y_\tau\ k(x_\sigma-x)^{k-1}k(y_\tau)^{k-1}(1-y_\tau-x_\sigma+x_\sigma y_\tau)^{L-2k-1} \\ &= \left[\frac{L!(L-2k-1)!}{(L-k)!(L-k-1)!}\right]^2(1-x)^{L-k-1}, \end{aligned} \quad (7.23)$$

so that

$$\lim_{L\to\infty} \mathbb{E}^{\frac{X}{L}}\left[\frac{\tilde\theta_k}{L}\right] = e^{-X}. \quad (7.24)$$

Finally, $\tilde\theta_k/L - \theta_k/L$ is a non-negative random variable with an expectation going to zero; it thus converges to zero in probability. Therefore, in the $L \to \infty$ limit by Slutsky's theorem, $\tilde\theta_k/L$ and θ_k/L have the same distribution as soon as one of the limits exists.

It now simply remains to notice that

$$\frac{\tilde\theta_k}{L} = \Big(\sum_{|\sigma|=k} n_\sigma(1-x_\sigma)^{L-2k-1}\Big)\Big(\sum_{|\tau|=L-k} m_\tau(1-y_\tau)^{L-2k-1}\Big), \quad (7.25)$$

which means that $\tilde\theta_k/L$ can be written has a contribution coming from the k first steps of the hypercube times an independent contribution coming from the k last steps. The contribution from the start depends on the value x of the origin. By symmetry, the contribution from the end has the same law as the contribution from the start with $x = 0$.

7.3. **Third step: the start of the hypercube is like a tree.** We now focus on the first term in (7.25):

$$\phi_k = \sum_{|\sigma|=k} n_\sigma (1 - x_\sigma)^{L-2k-1}. \tag{7.26}$$

The goal is to show that for a starting point $x = X/L$, in the large L limit then in the large k limit, this ϕ_k converges weakly to e^{-X} times an exponential distribution. Our strategy is to compare ϕ_k (defined on the first k levels of the hypercube) to the θ_k/L of the tree by showing that in the $L \to \infty$ limit the two quantities have the same generating function.

The difficulty, of course, is that one cannot write directly a recursion on the generating function of ϕ_k as we did on the tree because the paths after the first step are not independent. To overcome this, we introduce another quantity $\tilde{\phi}_k(b)$ which is (in a sense) nearly equal to ϕ_k:

$$\tilde{\phi}_k(b) = \sum_{|\sigma|=k} \tilde{n}_\sigma(b)(1 - x_\sigma)^L, \tag{7.27}$$

where we will shortly explain the meaning of the parameter b and give the definition of $\tilde{n}_\sigma(b)$. For now, let us just say that $\tilde{n}_\sigma(b) \le n_\sigma$; in other words, we discard some open paths when computing $\tilde{\phi}_k(b)$. It is clear that

$$\tilde{\phi}_k(b) \le \phi_k \tag{7.28}$$

and we will choose $\tilde{n}_\sigma(b)$ in such a way that

$$\lim_{L\to\infty} \mathbb{E}^{\frac{X}{L}}\big[\tilde{\phi}_k(b)\big] = \lim_{L\to\infty} \mathbb{E}^{\frac{X}{L}}\big[\phi_k\big]. \tag{7.29}$$

With the same argument as before, (7.28) and (7.29) will be sufficient to conclude that if $\lim_{L\to\infty} \tilde{\phi}_k(b)$ exists (we will show it is the case), then $\lim_{L\to\infty} \phi_k$ exists as well and has the same distribution. Then, we will be able to write a recursion for the generating function of $\tilde{\phi}_k(b)$ and solve it in the $L \to \infty$ limit.

Before going further, let us recall the following standard representation of the hypercube: to each node of the hypercube, we associate a different binary word with L bits (digits) in such a way that the starting point is $(0, 0, \ldots, 0)$, the end point is $(1, 1, \ldots, 1)$ and making a step is changing a single zero into a one. A node σ at level k has a label with exactly k ones.

We can now define b and $\tilde{n}_\sigma(b)$. The parameter b is a set of forbidden bits. Any path going through any bit in b is automatically discarded. In other words, $\tilde{n}_\sigma(b) = 0$ if σ has any bit equal to 1 which is in b. The parameter $\tilde{n}_\sigma(b)$ is 1 or 0, depending on whether there is an "interesting" path or not to σ. An interesting path is defined recursively in the following way:

- From the origin, we consider which nodes amongst the $L-|b|$ reachable first level nodes have a value which is smaller than $(\ln L)/L$; these are the "interesting" nodes at first level, and only the paths going through these interesting nodes are deemed interesting and are counted in $\tilde{n}_\sigma$.
- Let b' be the bits corresponding to all the interesting nodes at first level. After the first step, these b' bits are now forbidden for all interesting paths.

- Given the forbidden bits, the region of the hypercube reachable from each interesting node at first level is a sub-hypercube of dimension $L-|b|-|b'|$. All these hypercubes are non-overlapping. The construction of the interesting paths from each first level interesting node is now done recursively in the same way on each corresponding sub-hypercube.

Notice that by construction $\tilde n_\sigma(b)=0$ if $x_\sigma>(\ln L)/L$. This is a small price to pay as we expect that only the x_σ of order $1/L$ contribute. Furthermore, at each step we exclude $\mathcal{O}(\ln L)$ bits. For each open paths, at step k, there will therefore be $k\mathcal{O}(\ln L)$ forbidden bits. This is very small compared to L and will become negligible in the large L limit.

The definition of $\tilde n_\sigma(b)$ leads directly to a recursion on $\tilde\phi_k(b)$:

$$\tilde\phi_k(b,\text{starting point}=x)=\sum_{\rho\in b'}\mathbb{1}(x\le x_\rho)\tilde\phi^{(\rho)}_{k-1}(b\cup b',\text{starting point}=x_\rho),\qquad(7.30)$$

where b' is the (random) set of interesting first level nodes, those with a value smaller than $(\ln L)/L$ which avoid the b forbidden bits. *Given* b', for each bit $\rho\in b'$, $\tilde\phi^{(\rho)}_{k-1}$ is an *independent* copy of the variable defined in (7.27) with a different starting point. The recursion is initialized by

$$\tilde\phi_0(b)=(1-x)^L,\qquad(7.31)$$

which is non-random and independent of b.

Before computing the expectation and the generating function, remark that the distribution of $\tilde\phi_k(b)$ depends only on the number $|b|$ of forbidden bits, not on the bits themselves. We will abuse this remark and consider from now on that in the expression $\mathbb{E}^x[\tilde\phi_k(b)]$, the parameter b is actually the *number* of forbidden bits.

Let us now compute the expectation of $\tilde\phi_k(b)$. The distribution of the number b' of interesting nodes is binomial and we call $p(b')$ its law:

$$p(b')=\binom{L-b}{b'}\left(\frac{\ln L}{L}\right)^{b'}\left(1-\frac{\ln L}{L}\right)^{L-b-b'}.\qquad(7.32)$$

Then from (7.30)

$$\mathbb{E}^x\big[\tilde\phi_k(b)\big]=\sum_{b'=0}^{L-b}p(b')\times b'\int_x^{\frac{\ln L}{L}}\frac{L\,\mathrm{d}y}{\ln L}\ \mathbb{E}^y\big[\tilde\phi_{k-1}(b+b')\big].\qquad(7.33)$$

We will show by recurrence that the dependence in b can be written as

$$\mathbb{E}^x\big[\tilde\phi_k(b)\big]=\frac{(L-b)!}{(L-b-k)!\ L^k}\psi_k(x,L).\qquad(7.34)$$

It is obvious from (7.31) that this works for $k=0$. Assume that it works at level $k-1$. Then

$$\mathbb{E}^x\big[\tilde\phi_k(b)\big]=\frac{1}{L^{k-1}}\sum_{b'=0}^{L-b}p(b')\frac{(L-b-b')!}{(L-b-b'-k+1)!}b'\int_x^{\frac{\ln L}{L}}\frac{L\,\mathrm{d}y}{\ln L}\psi_{k-1}(y,L).\qquad(7.35)$$

The sum on b' decouples from from the integral and can be computed; one finds

$$\sum_{b'=0}^{L-b} p(b') \frac{(L-b-b')!}{(L-b-b'-k+1)!} b' = \frac{(L-b)!}{(L-b-k)!} \frac{\ln L}{L} \left(1 - \frac{\ln L}{L}\right)^{k-1} \tag{7.36}$$

and one recovers (7.34) with

$$\psi_k(x, L) = \left(1 - \frac{\ln L}{L}\right)^{k-1} \int_x^{\frac{\ln L}{L}} L \, \mathrm{d}y \, \psi_{k-1}(y, L) \tag{7.37}$$

or

$$\psi_k\Big(\frac{X}{L}, L\Big) = \left(1 - \frac{\ln L}{L}\right)^{k-1} \int_X^{\ln L} \mathrm{d}Y \, \psi_{k-1}\Big(\frac{Y}{L}, L\Big). \tag{7.38}$$

From here and $\psi_0(x, L) = (1-x)^L$, it is straightforward to show by recurrence that

$$\psi_k(X/L, L) \le e^{-X}. \tag{7.39}$$

Then, with this bound and the dominated convergence theorem, the limit of the integral in (7.38) is the integral of the limit and one shows by another straightforward recurrence that $\lim_{L\to\infty} \psi_k(X/L, L) = e^{-X}$.

Going back to (7.34), one then gets for any function $b(L)$ such that $b(L) = o(L)$

$$\mathbb{E}^{\frac{X}{L}}\big[\tilde{\phi}_k(b)\big] \le e^{-X}, \qquad \lim_{L\to\infty} \mathbb{E}^{\frac{X}{L}}\big[\tilde{\phi}_k(b(L))\big] = e^{-X}. \tag{7.40}$$

This completes the proof that $\tilde{\phi}_k(b)$ and ϕ_k have the same distribution in the $L \to \infty$ limit if that limit exists.

We now compute the distribution of $\tilde{\phi}_k(b)$ by writing a generating function. For $\mu \ge 0$, let

$$G_k(\mu, x, L, b) = \mathbb{E}^x\big[\exp\big(-\mu\tilde{\phi}_k(b)\big)\big]. \tag{7.41}$$

(Here again, we consider that the parameter b of G_k is a number.) From (7.30),

$$\begin{aligned} G_k(\mu, x, L, b) &= \sum_{b'=0}^{L-b} p(b') \left[\frac{L}{\ln L}\left(x + \int_x^{\frac{\ln L}{L}} \mathrm{d}y \, G_{k-1}(\mu, y, L, b+b')\right)\right]^{b'} \\ &= \sum_{b'=0}^{L-b} p(b') \left[1 - \frac{L}{\ln L} \int_x^{\frac{\ln L}{L}} \mathrm{d}y \, \Big[1 - G_{k-1}(\mu, y, L, b+b')\Big]\right]^{b'}. \end{aligned} \tag{7.42}$$

So

$$G_k\Big(\mu, \frac{X}{L}, L, b\Big) = \sum_{b'=0}^{L-b} p(b') \left[1 - \frac{1}{\ln L} \int_X^{\ln L} \mathrm{d}Y \, \left[1 - G_{k-1}\Big(\mu, \frac{Y}{L}, L, b+b'\Big)\right]\right]^{b'}. \tag{7.43}$$

If the $G_{k-1}(\ldots)$ on the right hand side did not depend on b', one could compute exactly the sum on b'. We will write bounds on G_{k-1} using quantities that do not depend on b' and compute this sum.

To do this, remark that G_k is an increasing function of b. Indeed, as we forbid more bits (b increases), we close more open paths, $\tilde{\phi}_k(b)$ decreases (or remains constant) and, from (7.41), G_k increases.

Therefore, a lower bound is easy: $G_{k-1}(\mu, Y/L, L, b+b') \geq G_{k-1}(\mu, Y/L, L, b)$ and

$$G_k\Big(\mu, \frac{X}{L}, L, b\Big) \geq \left[1 - \frac{1}{L}\int_X^{\ln L} \mathrm{d}Y \,\left[1 - G_{k-1}\Big(\mu, \frac{Y}{L}, L, b\Big)\right]\right]^{L-b}. \tag{7.44}$$

To obtain an upper bound, we use the fact that according to p, the probability that b' is larger than $\ln^2 L$ is very small. Then, in (7.43), we cut the sum over b' into two contributions. In the first part b' runs from 0 to $\lfloor \ln^2 L \rfloor$ and in the second part it runs from $\lfloor \ln^2 L \rfloor + 1$ to $L-b$. In the first part, we write $G_{k-1}(\mu, Y/L, L, b+b') \leq G_{k-1}(\mu, Y/L, L, b + \lfloor \ln^2 L \rfloor)$ and extend again the sum to $L-b$. In the second part we write that the term multiplying $p(b')$ is smaller than 1. Hence

$$\begin{aligned} G_k\Big(\mu, \frac{X}{L}, L, b\Big) \leq {}& \left[1 - \frac{1}{L}\int_X^{\ln L} \mathrm{d}Y \,\left[1 - G_{k-1}\Big(\mu, \frac{Y}{L}, L, b + \lfloor \ln^2 L \rfloor\Big)\right]\right]^{L-b} \\ &+ \sum_{b'=\lfloor \ln^2 L \rfloor + 1}^{L-b} p(b'). \end{aligned} \tag{7.45}$$

The remaining sum is of course the probability that b' is larger than $\ln^2 L$, which is vanishingly small as b' is binomial of average and of variance smaller than $\ln L$.

We can now show that $G_k(\mu, X/L, L, b)$ has a large L limit by recurrence. More precisely, we will show that for any function $b(L)$ which is a $o(L)$,

$$\widetilde{G}_k(\mu, X) := \lim_{L\to\infty} G_k\Big(\mu, \frac{X}{L}, L, b(L)\Big) \tag{7.46}$$

exists and is independent of $b(L)$.

This is obvious for $k=0$ as $G_0(\mu, x, L, b) = \exp[-\mu(1-x)^L]$, so that

$$\widetilde{G}_0(\mu, X) = \exp\big[-\mu e^{-X}\big]. \tag{7.47}$$

Suppose that (7.46) holds up to level $k-1$. Then for any function $b(L) = o(L)$, the function $b(L) + \lfloor \ln^2 L \rfloor$ is also an $o(L)$. We know from (7.41) and (7.40) that $G_k(\mu, X/L, L, b) \geq 1 - \mu \mathbb{E}^{\frac{X}{L}}[\tilde{\phi}_k(b)] \geq 1 - \mu e^{-X}$, so that we can use the dominated convergence theorem and obtain

$$\lim_{L\to\infty} \int_X^{\ln L} \mathrm{d}Y \,\left[1 - G_{k-1}\Big(\mu, \frac{Y}{L}, L, o(L)\Big)\right] = \int_X^{\infty} \mathrm{d}Y\, \big[1 - \widetilde{G}_{k-1}(\mu, Y)\big]. \tag{7.48}$$

It is then straightforward from (7.44) and (7.45) to see that (7.46) holds at level k and that

$$\widetilde{G}_k(\mu, X) = \exp\left[-\int_X^{\infty} \mathrm{d}Y\, \big[1 - \widetilde{G}_{k-1}(\mu, Y)\big]\right]. \tag{7.49}$$

Equations (7.47) and (7.49) are the same as (4.15), which completes the proof.

References

[1] Franke, J., Klözer, A., de Visser, J.A.G.M. and Krug, J. (2011). Evolutionary accessibility of mutational pathways, *PLoS Comput. Biol.* **7**, e1002134, 9pp.

[2] Hegarty, P. and Martinsson, A. (2012+). On the existence of accessible paths in various models of fitness landscapes. `ArXiv 1210.4798` [math.PR]

[3] Kauffmann, S. and Levin, S. (1987). Towards a general theory of adaptive walks on rugged landscapes. *J. Theoret. Biol.* **128**, 11–45.

[4] Kingman, J.F.C. (1978). A simple model for the balance between selection and mutation. *J. Appl. Probab.* **15**, 1–12.

[5] Lalley, S.P. and Sellke, T. (1987). A conditional limit theorem for frontier of a branching Brownian motion. *Ann. Probab.* **15**, 1052–1061.

[6] Nowak, S. and Krug, J. (2013+) Accessibility percolation on n-trees. `ArXiv 1302.1423` [cond-mat.stat-mech]

[7] Weinreich, D.M., Delaney, N.F., DePristo, M.A. and Hartl, D.M. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114.

[8] Weinreich, D.M., Watson, R.A. and Chao, L. (2005). Perspective: Sign epistasis and genetic constraints on evolutionary trajectories. *Evolution* **59**, 1165–1174.

[9] *On-line Encyclopedia of Integer Sequences*, **A003319**, `http://oeis.org/A003319`

Laboratoire de Probabilités et Modèles Aléatoires, UPMC, Paris, France
E-mail address: `julien.berestycki@upmc.fr`

Laboratoire de Physique Statistique, ENS, UPMC, CNRS, Paris, France
E-mail address: `Eric.Brunet@lps.ens.fr`

Laboratoire de Probabilités et Modèles Aléatoires, UPMC, Paris, France
E-mail address: `zhan.shi@upmc.fr`